

# Audio-Visual Object Extraction using Graph Cuts

Anna Llagostera Casanovas\* and Pierre Vanderghelynst

Signal Processing Laboratory (LTS2)

Ecole Polytechnique Fédérale de Lausanne (EPFL)

Station 11, 1015 Lausanne, Switzerland

e-mail: anna.llagostera@eecs.qmul.ac.uk, pierre.vanderghelynst@epfl.ch

phone: +41 21 693 26 01, fax: +41 21 693 76 00

**Abstract**—We propose a novel method to automatically extract the audio-visual objects that are present in a scene. First, the synchrony between related events in audio and video channels is exploited to identify the possible locations of the sound sources. Video regions presenting a high coherence with the soundtrack are automatically labelled as being part of the audio-visual object. Next, a graph cut segmentation procedure is used to extract the entire object. The proposed segmentation approach includes a novel term that keeps together pixels in regions with high audio-visual synchrony. When longer sequences are analyzed, video signals are divided into groups of frames which are processed sequentially and propagate the information about the source characteristics forward in time. Results show that our method is able to discriminate between audio-visual sources and distracting moving objects and to adapt within a short time delay when sources pass from active to inactive and vice versa.

**Index Terms**—Audio-visual processing, graph cut segmentation, synchrony, audio-visual object.

## I. INTRODUCTION

Humans combine audio and video modalities in a natural way. We can easily understand the relationship between an object that is falling and the sound of the crash, we intuitively link moving lips to the presence of speech, and we know what kind of music we will hear when we see a musical instrument being played. The fusion of the information perceived by both senses allows us better understand a scene than when considering each modality separately. Researchers have been trying to emulate the human behavior by performing a joint processing of audio and video signals for several applications. Nowadays, the video signal can be used to improve results in the audio domain for applications such as speech recognition, speech enhancement and sound source separation [1–10]. The coherence between audio and video modalities is also used to track or locate sound sources in the video signal [11–18]. Other approaches try to separate the scene into audio-visual sources, which are composed by a set of video structures and the associated sounds [19–21]. In a more general way, these applications can be used for automatic management of videoconferences, indexing and segmentation of multimedia data, video surveillance and robotics [22–24].

This work was supported by the Swiss NFS through grant 200021-117884 and by the EU Framework 7 FET-Open project FP7-ICT-225913-SMALL.

\*Anna Llagostera Casanovas contributed to this work while at the Signal Processing Laboratory (LTS2), EPFL, Switzerland. Now, she is with the with the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK.

Approaches in the joint audio-visual domain are based in a main assumption: related events in audio and video channels happen approximately at the same time. They follow similar strategies to assess the *synchrony* between both modalities. First, they define features for each modality such as the energy [17, 18, 25] or Mel-Frequency Cepstral Coefficients (MFCC) [12–14] for the audio, and pixel intensities [16, 25] or temporal variations [13, 14, 18] for the video. A fusion step combines then these representations by means of canonical correlation analysis [12, 18] or through the estimation of the joint densities of audio and video features [14, 15, 25]. Other approaches evaluate the synchrony between audio and video *structures* by decomposing in a first stage the two modalities over redundant dictionaries of signals, either separately [20] or jointly [26].

Even though a lot of effort has been devoted to the sound source localization task [12–18, 26], only three methods have attempted the extraction of the source’s video part [20, 27, 28]. The extracted speaker face can be used for example to protect the speaker’s identity or to emphasize him/her by blurring other persons and the background. The method in [20] decomposes the video signal into a set of image structures (atoms), and reconstructs the sources by clustering together atoms with high audio-visual correlation. Thus, in [20] the particular shapes of the sources are not considered, i.e. the extracted sources have always an approximately circular shape because all atoms inside a radius are used in the source reconstruction process. In [27, 28] the authors overcome this limitation by using a segmentation technique based on graph cuts, which is initialized by joint audio-visual analysis. In [27] the source position is estimated by computing the Quadratic Mutual Information between audio and video features, and this procedure is applied to sequences composed of almost static speakers. Then, in [28] this method is generalized to non-stationary sound sources by identifying the pixel’s visual trajectories whose changes in acceleration better fit the audio energy variations.

The method that we propose can also deal with non-stationary sound sources. First, the synchrony between audio and video channels is assessed, and regions moving coherently with the soundtrack are assigned to the audio-visual object. Next, a novel graph cut segmentation procedure is used to extract the entire audio-visual object. Longer sequences are analyzed by processing Groups of Frames (GoF) sequentially while transferring the obtained knowledge from GoF to GoF.

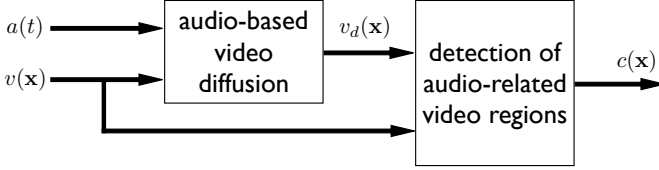


Fig. 1. Block diagram of the proposed approach for the identification of audio-related video regions. First, audio and video signals,  $a(t)$  and  $v(\mathbf{x})$ , are combined in a diffusion procedure that reduces the information in video regions which are not associated to the soundtrack. A second step compares the original video signal  $v(\mathbf{x})$  to the diffused signal  $v_d(\mathbf{x})$  to identify the possible location of audio-related video regions. The resulting signal, i.e. the audio-visual coherence  $c(\mathbf{x})$ , is high in video regions that have a high probability of belonging to an audio-visual object.

The main contributions of our approach are:

1. Our method extracts the audio-visual objects from *entire* sequences. Unlike previous approaches in this domain [27, 28], which only considered short video signals, our method can efficiently deal with longer sequences by propagating the segmentation results forward in time.
2. As a result, our method can deal with multiple audio-visual objects with different activity patterns. The extracted region evolves accordingly to the dynamics of the scene, i.e. each video region associated to a sound source is extracted only when sounds are generated by this source.
3. We propose a novel *audio-visual term* in the energy function that the graph cut algorithm minimizes. This term links together neighboring pixels in regions presenting a high correlation with the soundtrack and thus probably belonging to the audio-visual object. Unlike previous methods [27, 28], our term does not link regions presenting low audio-visual synchrony to the background. In consequence, the audio-visual object can be completely extracted even though some parts of it present a lower coherence.
4. We redefine the standard *regional term* in the energy function of the segmentation, which links each pixel to the foreground or background according to its color. Sec. IV illustrates the advantages of the proposed regional term over the commonly adopted term in [29–31]. Keeping this term represents also a significant advantage over the methods in [27, 28], since it ensures the cohesion between the homogeneous regions composing the audio-visual object.
5. The starting point of the segmentation process is determined automatically by means of joint audio-video analysis. The necessity of user interaction is the main limit of previous segmentation approaches [29–32].

This paper is structured as follows. Sec. II introduces a method to quantify the synchrony between image structures and sounds at the pixel level. Sec. III describes the automatic criterion for the choice of segmentation priors using audio-visual coherence. Sec. IV presents the proposed approach for the segmentation of audio-visual objects in a GoF. Sec. V explains the methodology that extracts the audio-visual objects from an entire sequence. Results are presented in Sec. VI and conclusions are drawn in Sec. VII.

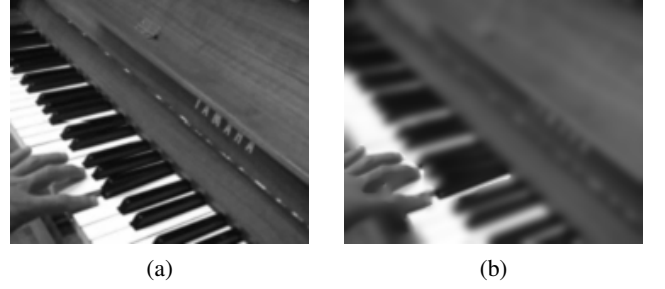


Fig. 2. Original frame (a) and resulting frame (b) after applying the audio-based video diffusion to a sequence in which a hand (sound source) plays a piano.

## II. IDENTIFICATION OF AUDIO-RELATED VIDEO REGIONS

The synchrony between related events in audio and video channels has been extensively exploited for the identification of video regions associated to the soundtrack. This research field was motivated by the presence of several studies which demonstrated the relationship between sounds and motion in the speech case [33–35]. Analogously to previous methods in this domain, the synchrony between audio and video modalities is used in our approach to identify the audio-visual objects in the scene. A block diagram of the proposed method is shown in Fig. 1.

In a first step, the synchrony between events in audio and video signals is assessed by using the audio-visual diffusion process in [36]. An event in the audio channel is defined as the presence of a sound, i.e. audio energy in the soundtrack. In the video channel an event is defined as the presence of some motion in the frame, which is estimated as the pixels inter-frame variation. Then, the diffusion procedure reduces the information (spatio-temporal edges) in video regions whose motion is not synchronous with the presence of sounds. Fig. 2 shows an example of a video signal before (a) and after (b) the audio-visual diffusion procedure. The hand in this scene (audio-visual object) remains well-defined and is better preserved from diffusion than the rest of the frame (e.g. piano keys) which is more blurred.

In the second step (right part in Fig. 1), regions in which the video signal is least diffused are identified by comparing the motion (temporal edges) before and after the audio-visual diffusion process. Regions in which the edges are well preserved are, with high probability, part of the audio-visual object since their movements are synchronous with the sounds.

Let  $v(\mathbf{x})$  be the video signal  $v$  at spatio-temporal coordinates  $\mathbf{x} = (x, y, t)$ , and  $v_d(\mathbf{x})$  be the resulting video signal after the diffusion procedure. Then, the *audio-visual coherence*  $c(\mathbf{x}) \in [0, 1]$  at pixel location  $\mathbf{x}$  is defined as

$$c(\mathbf{x}) = \begin{cases} \frac{1}{s} \frac{\partial_t v_d(\mathbf{x})}{\partial_t v(\mathbf{x})} & \text{if } \partial_t v(\mathbf{x}) > \xi \\ \frac{1}{s \arg\max_{\mathbf{x}} \partial_t v(\mathbf{x})} & \text{otherwise} \end{cases} \quad (1)$$

where  $\partial_t(\cdot)$  represents the derivative with respect to the time axis  $t$ , the constant  $\xi$  makes the audio-visual coherence close to zero in static pixels, and the constant  $s$  makes  $c(\mathbf{x})$  unitary. In this expression, the temporal derivative of the video signals before and after diffusion is approximated using finite differ-

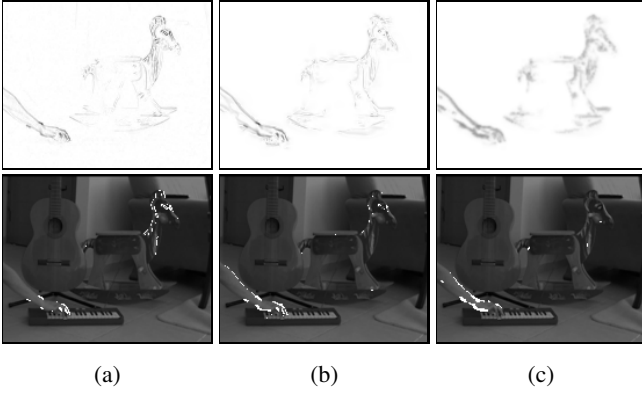


Fig. 3. White pixels in the bottom row indicate the 0.5% highest values of the top features: (a) original motion  $\partial_t v(\mathbf{x})$ , (b) resulting motion  $\partial_t v_d(\mathbf{x})$  after diffusion and (c) audio-visual coherence  $c(\mathbf{x})$ . In this clip a hand plays a piano while a rocking horse is moving. Darker regions in the top row indicate higher values for the features.

ences as the variation between pixels in consecutive frames. The higher is the audio-visual coherence  $c(\mathbf{x})$  the higher is the probability for the video pixel at location  $\mathbf{x}$  to be part of an audio-visual object, since the edges in this location are well preserved through the diffusion process.

Fig. 3 shows a frame of a sequence where a hand plays a piano. In this clip, the video motion generated by the audio-visual object has similar magnitude to the distracting motion, i.e. the highest values of the original motion in (a) are equally distributed between hand and horse. After the diffusion process the motion is more intense in the hand region (b). Finally, the audio-visual coherence in (c) is clearly dominant in the audio-visual object: the hand's silhouette is darker [top] and only a few white pixels appear over the rocking horse.

The *audio-visual coherence* is an efficient measure of the relationship between image structures and the sounds, with a high spatial resolution. This measure is used in Sec. III to automatically determine the starting point for the segmentation procedure, and in Sec. IV in the definition of the audio-visual term in the energy function of the segmentation.

Other measures could be used for the identification of the audio-related video regions. The proposed approach for the extraction of audio-visual objects is independent of the audio-visual synchrony measure that is used.

### III. AUDIO-VISUAL SEGMENTATION PRIORS

The extraction of the audio-visual object requires a starting point for the segmentation process, i.e. some initial information about the foreground (audio-visual object) and background location. In our method, this prior information (segmentation seeds) is obtained from the fusion of audio and video modalities. Pixels with high audio-visual coherence are likely to compose the audio-visual object because they belong to an image region moving synchronously with the sounds.

The foreground seeds are chosen to be the  $N_f$  pixels with highest audio-visual coherence  $c_p$ , while  $N_b$  background seeds are *randomly* distributed in the GoF. The random selection of the background seeds ensures that no additional assumptions are made. The number of seeds that are automatically chosen

for foreground  $N_f$  and background  $N_b$  are

$$N_m = P \cdot H_{AV} \quad \text{for } m = \{f, b\}, \quad (2)$$

where  $P$  is the number of pixels in the video GoF and the parameter  $H_{AV}$  determines the density of the seeds.

No segmentation seeds are fixed in the video frames in silent periods, because no knowledge about the sources can be extracted from the joint audio-visual processing (there are no active sources). Furthermore, since in this frames the audio-visual coherence is very low, the lack of foreground seeds combined with the introduction of background seeds would penalize the extraction of the audio-visual object.

### IV. GRAPH CUT SEGMENTATION USING AUDIO-VIDEO SYNCHRONY

Significant progress has been made in the last 20 years in the user-guided foreground/background segmentation domain. Among all segmentation techniques (snakes, active contours, shortest path techniques...) graph cuts have shown applicability to N-dimensional problems and flexibility in the definition of the energy to minimize. Furthermore, they provide a globally optimal segmentation through a numerically robust minimization procedure. Graph cuts were first introduced by Boykov and Jolly in [29] for the segmentation of monochrome N-D signals and extended to color images and videos in latter approaches [30–32]. For a detailed introduction to the recent advances in image and video segmentation, please refer to the report in [37].

The proposed 3D segmentation approach is inspired by the method in [29]. Given some initial information about foreground and background locations provided by the user (seeds) their algorithm computes a globally optimal segmentation using graph cuts. Our method integrates information extracted from joint audio-visual processing in the segmentation procedure. A preliminary work on this subject can be found in [38].

#### A. Formulation

Let  $\mathbf{z} = (z_1, \dots, z_p, \dots, z_P)$  be the set of  $P$  pixels in the RGB color space that compose a GoF. The segmentation task consists on assigning a binary label  $l = (l_1, \dots, l_P)$  to each pixel  $p$  in the GoF:  $l_p \in \{0(\text{background}), 1(\text{foreground})\}$ .

First, we build a graph  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$  corresponding to the GoF following the procedure in [29]. The set of vertices  $\mathcal{V}$  is composed of the pixels  $p \in \mathcal{P}^j$  in GoF  $j$  plus two additional nodes: a foreground terminal  $F$  and a background terminal  $B$ . The set of edges  $\mathcal{E}$  is composed by edges connecting neighboring pixels  $\{p, q\} \in \mathcal{N}$  (n-links) and edges connecting each pixel  $p$  to the foreground and background terminals  $\{p, F\}$  and  $\{p, B\}$  (t-links). In our graph the neighborhood  $\mathcal{N}$  of each pixel is composed of six pixels, four spatial neighbors and two temporal neighbors.

Then, the graph cut algorithm solves the segmentation problem by minimizing the following energy defined on the

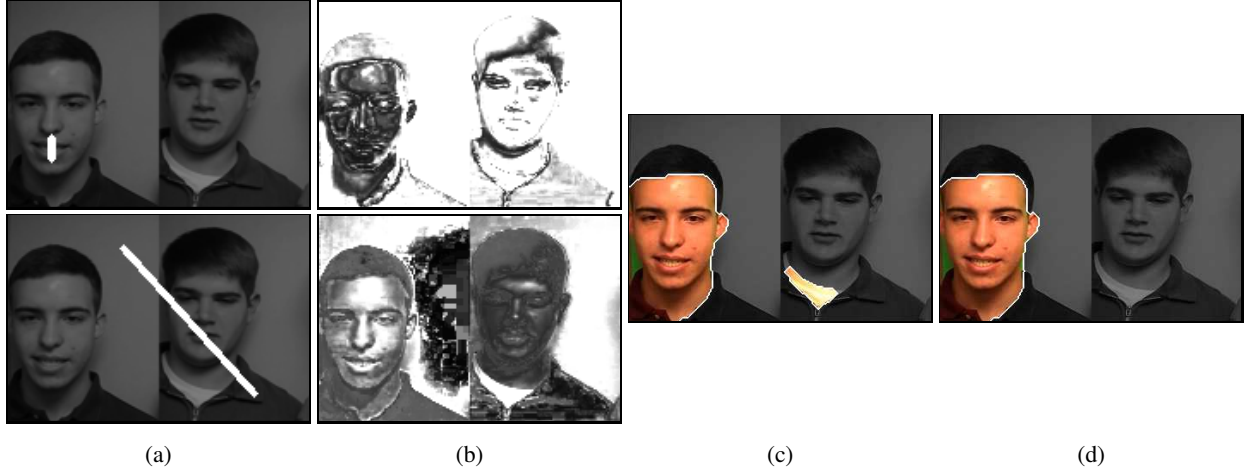


Fig. 4. Segmentation results when using the regional term in previous methods (c) and our regional term (d), given the manually-added seeds (a) and the corresponding probability maps (b) for foreground [top] and background [bottom]. The audio-visual term is not considered ( $\lambda_C = 0$ ). The extracted region is shown in color and the background in a darker grayscale. White regions represent the seeds in (a) and a very low probability in (b).

graph:

$$\begin{aligned} J(l) &= \lambda_R R(l) + V(l) + \lambda_C C(l) \\ &= \lambda_R \sum_{p \in \mathcal{P}^j} R_p(l_p) + \sum_{\{p,q\} \in \mathcal{N}} (V_{p,q} + \lambda_C C_{p,q}) [l_p \neq l_q], \end{aligned} \quad (3)$$

where  $[\Phi]$  denotes the indicator function taking values 0, 1 for a predicate  $\Phi$ . The regional term  $R(l)$  evaluates how the color  $z_p$  corresponding to the pixel  $p$  with label  $l_p$  fits into the background and foreground models, the boundary term  $V(l)$  assesses the similarity of each pixel with its neighborhood, and the audio-visual term  $C(l)$  links together neighboring pixels in regions presenting a high audio-visual coherence. The relative importance of the regional term and the audio-visual term with respect to the boundary term is determined by the coefficients  $\lambda_R$  and  $\lambda_C$ .

In our method, the energy  $J(l)$  is minimized using the Boost Graph Library implementation [39] of the classical minimum cut algorithm in [29].

### B. Boundary Term

The boundary term  $V(l)$  keeps together neighboring pixels with similar color. As in [29–31], our *boundary term* is defined by

$$V_{p,q} = \frac{1}{\text{dist}(p,q)} \exp\left(-\frac{\|z_p - z_q\|^2}{2\gamma_V^2}\right), \quad (4)$$

where  $\text{dist}(\cdot)$  is the Euclidean distance between neighboring pixels, both in space and time. We fix  $\gamma_V^2 = \mathbb{E}(\|z_p - z_q\|^2)$  as in [30], where  $\mathbb{E}(\cdot)$  denotes the expectation operator over the video signal.  $V_{p,q}$  is low when pixels  $p$  and  $q$  have significantly different colors, i.e.  $\|z_p - z_q\| > \gamma_V$ , and it is high when their colors are similar.

### C. Regional Term

The regional term  $R(l)$  evaluates the pixels similarity to the foreground and background color distributions. Our regional term is slightly different from previous methods [29–31].

First, foreground ( $\Lambda^f$ ) and background ( $\Lambda^b$ ) Gaussian Mixture Models (GMMs) are estimated using the Expectation Maximization algorithm on the available seeds:  $\Lambda^m = \{u_i^m, \mu_i^m, \Sigma_i^m\}_{i=1}^Q$  for  $m = \{b, f\}$ . For each Gaussian  $i$  composing the mixture,  $u_i$ ,  $\mu_i$  and  $\Sigma_i$  denote respectively its weight, mean and covariance matrix. The number of Gaussians is fixed to  $Q = 5$  in all experiments as in [30].

According to these color models, the penalties for assigning pixel  $p$  to foreground ( $l_p = 1$ ) and background ( $l_p = 0$ ) that compose the *regional term* are defined respectively as

$$\begin{aligned} R_p(l_p = 1) &= h(\ln P(z_p | \Lambda^b)), \\ R_p(l_p = 0) &= h(\ln P(z_p | \Lambda^f)), \end{aligned} \quad (5)$$

where  $P(z_p | \Lambda^m)$  is the probability for a pixel  $p$  to belong to the foreground/background given the color model  $\Lambda^m$ , and  $h(\cdot)$  is a function that maps  $\ln P(z_p | \Lambda^m)$  from  $(-\infty, 0]$  to  $[0, 1]$  where “0” and “1” represent the lowest and the highest probability respectively.

The weight of the edge that links a pixel  $p$  to the foreground (background) is proportional to the probability for its color  $z_p$  of belonging to the foreground (background) color model expressed by  $\Lambda^f$  ( $\Lambda^b$ ). Previous methods [29–31] used the negative log-likelihoods, and therefore the edge’s weight was *inversely* proportional to this probability. Fig. 4 illustrates the advantages of the proposed regional term. In this example, the segmentation seeds for foreground [top] and background [bottom] are manually selected as depicted in (a) and the audio-visual term is not taken into account. The probability maps according to the manually-selected seeds are shown in (b). The probability for a pixel situated in the right person’s shirt of belonging to both foreground and background is very low (in white in the central figures). According to the proposed regional term, the links between those pixels and the background and foreground terminals have a very low weight and therefore they do not influence the segmentation results. However, when using the term in [29–31] the link between the pixels in the shirt and the foreground terminal is much stronger than the link to the background because the

probability of belonging to the background is lower. Notice that the segmentation result contains the right person's shirt when applying the regional term in [29–31] (c), while it is not extracted in our case (d). The regional term in previous methods enforced the segmentation algorithm to label those pixels as foreground, even though this is not clear at all according to the color models. In our approach, we prefer to rely on the boundary term and do not influence the segmentation when the probabilities of belonging to foreground and background are so remote.

#### D. Audio-Visual Term

The proposed term keeps together pixels in regions that move in synchrony with the sounds. The *audio-visual term* it is defined by

$$C_{p,q} = \frac{1}{\text{dist}(p,q)} c_p \exp\left(-\frac{|c_p - c_q|^2}{2\gamma_C^2}\right), \quad (6)$$

where  $c_p$  is the audio-visual coherence  $c(\mathbf{x})$  corresponding to pixel  $p$  with spatio-temporal coordinates  $\mathbf{x}$ . Since in this case  $C_{p,q} \neq C_{q,p}$  if  $c_p \neq c_q$ , our graph is directed. The constant  $\gamma_C$  has a similar purpose than  $\gamma_V$  in the boundary term.  $C_{p,q}$  is low when pixels  $p$  and  $q$  have a significantly different coherence, i.e.  $|c_p - c_q| > \gamma_C$ , and  $C_{p,q} \approx c_p$  when the audio-visual coherence of the two pixels is similar.

Our audio-visual term is similar to the boundary term in the sense that it is computed between neighboring pixels. Low weights are assigned to the edges that link pixels in different regions (in this case regions presenting high and low coherence instead of regions with significantly different color). Our audio-visual term does not affect regions with low audio-visual coherence. The weight  $C_{p,q}$  is directly proportional to the audio-visual coherence in the origin pixel  $c_p$  and thus the weight of the links is close to zero in regions with low coherence. Therefore, the proposed audio-visual term only links together neighboring points that present a similar and *relevant* audio-visual coherence.

Two approaches had also introduced an audio-visual energy term in the segmentation process [27, 28]. First, audio-visual correlation values were clustered in two groups representing the sound source and the background. Then, the regional term,  $R(l)$  in Eq. (3), was *replaced* by a cost to assign a pixel to the sound source, which depended on the Mahalanobis distance between the pixel and the estimated mean value of the source's correlation. In contrast, here we keep the regional term and we add a novel audio-visual term. Our term links together neighboring pixels in regions with high audio-visual coherence instead of linking each pixel to the foreground and background terminals. Therefore, in our approach pixels composing the audio-visual object are kept together in the segmentation process, without affecting regions with low coherence (they were assumed to belong to the background in [27, 28]). Since the connections between pixels are spatio-temporal, our audio-visual term reinforces also the links between neighboring frames in regions where the image structures move in synchrony with the sounds.

TABLE I  
PROPOSED WEIGHT DISTRIBUTION FOR THE GRAPH.

edge	weight	for
$\{p, q\}$	$V_{p,q} + \lambda_C C_{p,q}$	$p, q \in \mathcal{N}$
$\{p, F\}$	$\lambda_R \cdot h(\ln P(z_p   \Lambda_f))$	$p \in \mathcal{P}^j, p \notin \mathcal{F}^j \cup \mathcal{B}^j$
	$L$	$p \in \mathcal{F}^j$
	$0$	$p \in \mathcal{B}^j$
$\{p, B\}$	$\lambda_R \cdot h(\ln P(z_p   \Lambda_b))$	$p \in \mathcal{P}^j, p \notin \mathcal{F}^j \cup \mathcal{B}^j$
	$0$	$p \in \mathcal{F}^j$
	$L$	$p \in \mathcal{B}^j$

#### E. Weight Summary

The distribution of the weights in the graph is summarized in Table I. Here  $p \in \mathcal{F}^j$  and  $p \in \mathcal{B}^j$  denote respectively the set of points in  $j$ -th GoF that are classified into foreground and background by the joint audio-visual analysis in Sec. III (segmentation seeds).

In general,  $L = 1 + \max_{p \in \mathcal{P}^j} \sum_{q: \{p,q\} \in \mathcal{N}} (V_{p,q} + \lambda_C C_{p,q})$  when the seeds are manually fixed, to ensure that the seeds label is not modified [29]. However, since the seeds choice is unsupervised in our approach, we fix the weight that links the seeds to the corresponding terminal ( $F$  or  $B$ ) to the maximum weight of a n-link:  $L = \max_{p \in \mathcal{P}} (V_{p,q} + \lambda_C C_{p,q})$ . This value is high enough to influence the segmentation but the initial label of the seeds can be modified by the minimum cut algorithm [29] if required, e.g. when a foreground seed is isolated in a region labelled as background.

### V. AUDIO-VISUAL OBJECT EXTRACTION ON ENTIRE SEQUENCES

In practice, video signals can not be processed globally due to their size. They are usually divided into parts that will be analyzed separately. Then, the problem relies on efficiently sharing the information among the different parts. Algorithm 1 summarizes the proposed approach for the extraction of audio-visual objects on entire sequences. The idea is to process Groups of Frames (GoFs) sequentially: when sounds appear the first GoF is segmented as explained in Sec. IV, and then in the following GoFs we combine the knowledge extracted from the previous GoF (i.e. location and characteristics of the sources) and the joint audio-visual processing on the current GoF. Our approach exploits the temporal coherence between neighboring frames and ensures the continuity of the segmentation results. In consequence, the GoFs are processed separately *but not independently*.

#### A. Global Processing

First, we apply the the procedure described in Sec. II to compute the audio-visual coherence  $c_p$  for each pixel  $p$  from the audio and video signals. This information represents the starting point for the audio-visual segmentation approach in Sec. IV and it will be used in all stages of our method.

**Input:** Video signal  $v(\mathbf{x})$  and audio signal  $a(t)$   
**Output:** Segmented video with binary labels  $l$

**A.** Compute the audio-visual coherence  $c_p$  for each pixel  $p \in \mathcal{P}$  from the audio and video signals  $a(t)$  and  $v(\mathbf{x})$ .  
**B.** Partition the video signal into  $M$  fixed-size GoFs, each one composed of  $N_t$  frames. Neighboring pairs of GoFs share one frame:  $\mathcal{P}^j \cap \mathcal{P}^{j+1} \neq \emptyset$ . First GoF starts when sounds appear.  
**for** GoF  $j = 1$  **do**  
    1. Classify the  $N_f$  pixels with highest audio-visual coherence into the foreground ( $p \in \mathcal{F}^1$ ) and choose randomly  $N_b$  pixels as background seeds ( $p \in \mathcal{B}^1$ ).  
    2. Learn color models for foreground and background ( $\Lambda_{f,1}$  and  $\Lambda_{b,1}$ ) from the audio-visual seeds in this GoF.  
    3. Segment the GoF and obtain the labels  $l^1$  and corresponding trimap  $T^1$  given the color models  $\Lambda_{f,1}$ ,  $\Lambda_{b,1}$ , and seeds  $p \in \mathcal{F}^1$ ,  $p \in \mathcal{B}^1$ . The value of the trimap  $T_p$  at pixel  $p$  is 1 in the foreground, 0 in the background and 0.5 in the border between the two regions.  
**end**  
**for each** GoF  $j = 2, \dots, M$  **do**  
    1. Fix  $N_f$  and  $N_b$  *audio-visual seeds* ( $p \in \mathcal{F}^j$ ,  $p \in \mathcal{B}^j$ ) following the same procedure as for the first GoF.  
    2. Add  $N_f^c$  and  $N_b^c$  *continuity seeds* according to the segmentation result on the shared frame as  

$$p \in \mathcal{F}^j \leftarrow p \in \mathcal{F}^j \cup R_{N_f^c}\{\mathcal{C}_f\}, \quad (7)$$

$$p \in \mathcal{B}^j \leftarrow p \in \mathcal{B}^j \cup R_{N_b^c}\{\mathcal{C}_b\}. \quad (8)$$
  
 $R_N\{\psi\}$  denotes the restriction of the set  $\psi$  to  $N$  of its values chosen uniformly *at random*, and  $\mathcal{C}_f$ ,  $\mathcal{C}_b$  are the set of all possible pixels to use as continuity seeds, which are labelled as foreground and background in the trimap  $T^{j-1}$ :  

$$\mathcal{C}_f = \{p \in \{\mathcal{P}^{j-1} \cap \mathcal{P}^j\} : T_p^{j-1} = 1\}, \quad (9)$$

$$\mathcal{C}_b = \{p \in \{\mathcal{P}^{j-1} \cap \mathcal{P}^j\} : T_p^{j-1} = 0\}. \quad (10)$$
  
    3. Compute the color models ( $\Lambda_{f,j}$  and  $\Lambda_{b,j}$ ) using the audio-visual seeds in this GoF *and* continuity seeds in the shared frame.  
    4. Segment the GoF  $j$  and obtain the labels  $l^j$  and the corresponding trimap  $T^j$  given the color models  $\Lambda_{f,j}$ ,  $\Lambda_{b,j}$  and seeds  $p \in \mathcal{F}^j$ ,  $p \in \mathcal{B}^j$ .  
**end**

**Algorithm 1:** Audio-Visual Object Extraction

Next, the video signal is divided into fixed-size GoFs, each neighboring pair of GoFs sharing one frame. This configuration is chosen for two main reasons. First, all video GoFs have the same size and thus the same graph structure. As a result, the graph is built for the first GoF and then reused in the next ones (only the weights change)<sup>1</sup>. The second and most important reason for which we have chosen this GoF structure is that the frame that two neighboring GoFs share facilitates the propagation of the segmentation results. Indeed, the seeds that are used in the segmentation of a GoF are obtained from the audio-visual analysis in Sec. III *and* the segmentation results in the shared frame. Thus, this frame links neighboring GoFs and allows the introduction of prior information in the segmentation of the GoF.

The extraction of the audio-visual object starts when the first sounds are captured by the microphone, since an audio-visual object has, by definition, an audio part associated to it.

### B. First GoF Processing

The audio-visual object in the first GoF is extracted as explained in Secs. III and IV. First, seeds are chosen according to the reasoning in Sec. III: the  $N_f$  pixels presenting the highest audio-visual coherence become seeds for the foreground  $p \in \mathcal{F}^1$  and the same number ( $N_b = N_f$ ) of background seeds  $p \in \mathcal{B}^1$  are uniformly distributed at random across the GoF. Next, GMMs are estimated for the foreground and background color distributions on the available seeds, which are obtained by joint audio-visual processing. Finally, the segmentation is computed according to the procedure detailed in Sec. IV.

To avoid the propagation of errors, the limits of the segmentation in the shared frame are dilated and eroded to build a trimap  $T$  indicating locations where the labels have enough confidence. Fig. 5 shows an example of a segmented frame (a) and the corresponding trimap (b). The value of the trimap is 1 in the foreground (white), 0 in the background (black) and 0.5 in the border between the two regions (gray).

### C. Next GoFs Processing

We exploit the temporal consistency that characterizes video signals (neighboring frames are usually very similar). In fact, the characteristics of the audio-visual objects in the scene

<sup>1</sup>Another possibility could be to determine the GoF size by detecting some specific features such as scene changes. In this case, a possible improvement in the performance around the scene cut would come at the expense of more computational cost.



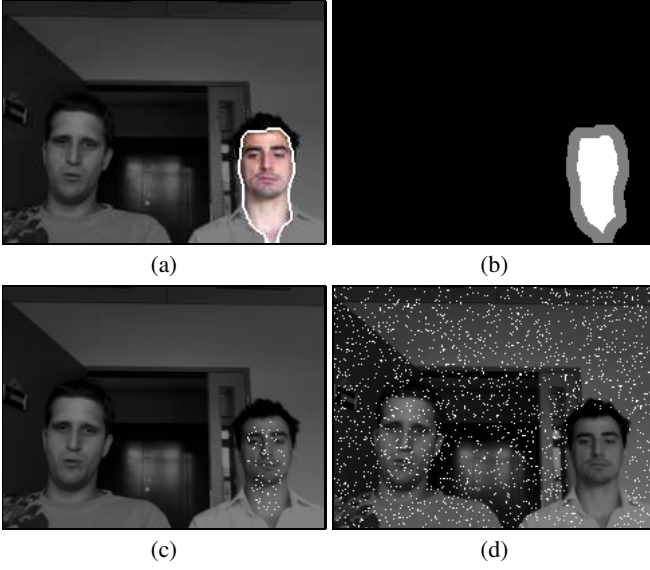


Fig. 5. Extraction of the *continuity seeds* in the first frame of an intermediate GoF. Segmentation result (a) obtained from the previous GoF's processing, corresponding trimap (b) when dilating and eroding the segmentation boundaries, and foreground (c) and background (d) continuity seeds (white pixels).

(e.g. position, shape and color statistics) do not change much from frame to frame *unless multiple sources with different activity patterns are present*. For this reason, we keep the same segmentation procedure than for the first GoF while adding some knowledge about the previous GoF, i.e. we add a continuity prior.

*Continuity seeds* are used to ensure the temporal consistency between GoFs. The continuity seeds are selected *randomly* from the set of pixels in the shared frame that are labelled as foreground and background in the trimap, i.e.  $\mathcal{C}_f$  and  $\mathcal{C}_b$  in Eq. (9)-(10) in Algorithm 1. The number of continuity seeds is determined by

$$N_m^c = |\mathcal{C}_m| H_C \quad \text{for } m = \{f, b\} \quad (11)$$

where  $|\mathcal{C}_m|$  denotes the cardinality of  $\mathcal{C}_m$  and the parameter  $H_C$  controls the density of the continuity seeds in the shared frame. The higher is  $H_C$ , the more continuity seeds we fix, and the more we rely on the prior information. If we decrease  $H_C$  we reduce the influence of segmentation result obtained for the previous GoF and  $H_C = 0$  is equivalent to processing each GoF independently. Fig. 5 (c)-(d) show examples of foreground and background continuity seeds when  $H_C = 0.05$ . The set of segmentation seeds in the GoF is thus composed of the continuity seeds in the first frame (shared frame) *and* the audio-visual seeds in the remaining frames, which are chosen as described in Sec. III.

In the first GoF the color models are learned on the audio-visual seeds. In the following GoFs, more information is available, since we know the color distributions of the audio-visual object and the background in the previous GoF. Our method uses both continuity seeds and audio-visual seeds in the estimation of the color models. When a new source becomes active, its colors are introduced in the foreground model  $\Lambda_{f,j}$  by means of the audio-visual seeds. In addition,



Fig. 6. Results when varying the proportion of audio-visual seeds and continuity seeds for the color models estimation on some frames from *Speakers2* where two sources alternate their periods of activity. (a) Only seeds from audio-visual processing are used in the GMM computation, (b) a 10% of continuity seeds are introduced, and (c) the continuity seeds represent a 50%. At the beginning only the right person is speaking (frames  $t_1$  to  $t_3$ ), then he stops and the left person starts speaking (frames  $t_4$  to  $t_7$ ).

if a source is active in two consecutive GoFs the borders of the segmented region will remain stable because the color distribution of the source is also introduced in the foreground model by means of the continuity seeds. The ratio of audio-visual seeds used in the estimation of the color models is determined by the parameter  $\beta$ . Fig. 6 shows the extracted audio-visual object for different values of  $\beta$ . In all cases the segmentation seeds include the audio-visual and continuity seeds, i.e. the percentages considered in this analysis only affect the estimation of  $\Lambda_{f,j}$  and  $\Lambda_{b,j}$ . Results show that the borders of the extracted object are very unstable when using only the audio-visual information in Fig. 6(a). In some frames, the audio-visual seeds are highly concentrated in the

TABLE II  
PARAMETERS SUMMARY.

$\xi$	$\lambda_R$	$\lambda_C$	$\gamma_C$	$H_{AV}$	$H_C$	$\beta$
0.1	0.05	0.6	0.1	0.003	0.05	0.9

mouth region, the foreground color model only captures the color distribution in this region and the rest of the face is not extracted (frames  $t_2$  and  $t_6$ ). In contrast, if the color models rely too much on the prior information the same region as in the previous GoF is extracted even if that source is not currently active. In Fig. 6(c) the right person is extracted for a considerably long period after becoming inactive (the time difference between  $t_3$  and  $t_5$  is around 2 seconds). A good compromise is reached in Fig. 6(b), where the continuity seeds ensure the stability of the audio-visual object borders, and the extracted region is able to switch easily between the active speakers.

Once the seeds and the color models are estimated, the GoF is segmented by applying the procedure described in Sec. IV. At each step of the algorithm the extracted region is dilated and eroded to obtain a new trimap and reduce the risk of propagating segmentation errors through time.

To summarize, there needs to be a balance between the amount of information that we use from the temporal consistency and from the audio-visual analysis. In our approach, the information extracted from joint audio-visual processing is prevailed over the knowledge about the active source in the former GoF, so that the extracted region is affected *but not determined* by the previous segmentation result.

## VI. EXPERIMENTS

This section is divided into two parts. Sec. VI-A presents the results when extracting audio-visual objects in fragments of sequences (i.e. one GoF), validating thus Secs. II to IV. Sec. VI-B shows the results obtained on entire video sequences and tests the entire scheme for the extraction of audio-visual sources explained in Sec. V. Our dataset is composed of clips belonging to the *groups* section of the CUAVE database [40], movies from two state-of-the-art source localization approaches [18, 41], and an additional sequence recorded in a realistic office environment to test complementary aspects of our approach.

The average processing time for automatically segmenting a video frame in a MacBook Pro laptop machine with an Intel Core 2 Duo CPU at 2.4 GHz and 2GB memory is about 2.5s: 1.6s for the selection of audio-visual priors and 0.9s for the graph cut segmentation procedure. However, the audio-visual diffusion process that is needed to compute the audio-visual coherence and determine the segmentation priors has not been optimized for the moment. It is currently coded in MATLAB and thus the processing time required for the choice of the segmentation seeds can drop drastically when parallelized. Notice that in the discrete formulation of the diffusion process in [36] the value of the video signal at each point only depends on its six spatio-temporal neighbors.

The main parameters in the proposed approach for the extraction of audio-visual objects are *fixed in all experiments* as shown in Table II.  $\lambda_R = 0.05$ , a value within the range defined by [31] and [30], and  $\lambda_C = 0.6$  so that the extracted region respects the strong edges in the image (the audio-visual term has a lower weight than the boundary term). However, results do not change significantly for  $\lambda_C \in [0.5, 1.5]$  and  $\lambda_R \in [0.04, 0.06]$ .  $\gamma_C = 0.1$  allows high values for the audio-visual term when two pixels have similar and high audio-visual coherence, because  $c(\mathbf{x})$  is unitary and therefore the maximum value of  $|c_p - c_q|$  in Eq. (6) is 1.  $H_{AV}$  is low to introduce the smallest possible number of errors in the segmentation priors, i.e. only a 0.3% of the pixels are selected as audio-visual seeds. The continuity seeds are composed by the 5% of pixels of the shared frame whose labels are clear according to the trimap ( $H_C = 0.05$ ), so that the result on the previous GoF does not determine the region to extract in the current GoF. Each GoF is composed of  $N_t = 20 - 25$  frames depending on the sampling rate of the analyzed sequence (GoFs are around 1 second long) for two main reasons. First, hardware restrictions make it difficult to segment long time intervals due to the large number of vertices in the graph. Second, we want  $N_t$  small to allow fast transitions in the extracted regions when a source switches from active to inactive or vice versa. Notice that it is difficult to extract an audio-visual object for only a part of the GoF, since the regional and boundary terms link together homogeneous regions with similar color statistics.

### A. Results on One Video GoF

First, results obtained with our method are compared to those reported in previous audio-visual segmentation approaches [27, 28]. Next, we demonstrate the importance that the proposed audio-visual term has on the segmentation result. Finally, we show that our method is able to extract multiple sources that are active at the same time.

Fig. 7 compares the extracted audio-visual objects obtained with our method [bottom] and the methods in [27, 28] [top] when analyzing several fragments of clips *g22* and *g23* of CUAVE database [40]. Our results are specially favorable in (c): the region that we extract contains the complete mouth region while in [27] it was mostly composed of the girl's hair. In (e) our approach extracts completely the girl's face because the presence of the regional term makes easier the extraction of regions homogeneous in color. The term in [27, 28] penalizes pixels presenting a low coherence with the soundtrack by linking them to the background, and consequently only the mouth region can be extracted in (e). Therefore, our method seems more suitable for applications that require the entire face region of the current speaker, or a more complete source region in general. An example of such an application could be the protection of the speaker's identity by automatically mosaicing his/her face.

Fig. 7 also compares the results *with* and *without* the audio-visual term in Eq. (3). When the audio-visual term is not considered ( $\lambda_C = 0$ ) the speaker's mouth region is only partially extracted in (a) and (c). The introduction of the proposed audio-visual term links together the pixels in the



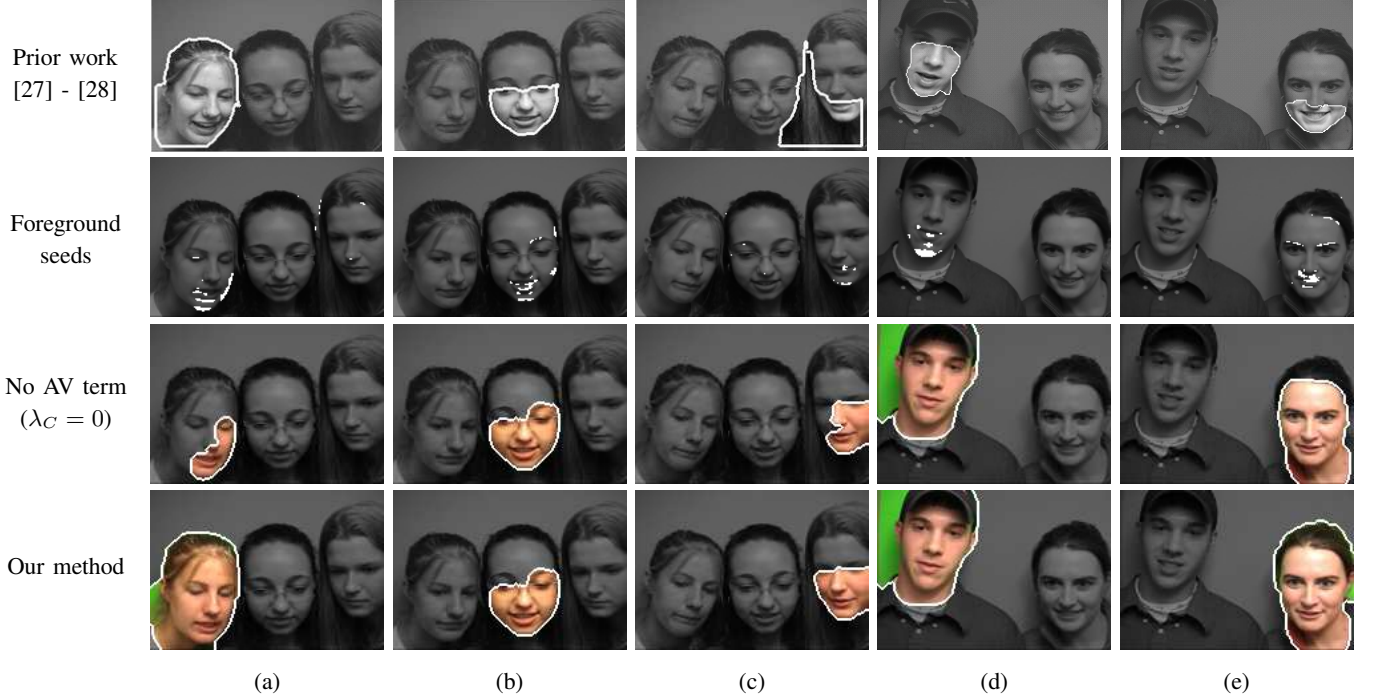


Fig. 7. Comparison between the results obtained with the proposed approach and the methods in [27, 28], and effect of the audio-visual term in the segmentation process. [From top to bottom] Extracted regions when applying the method in [27] to clip *g23* (a)-(c) and the approach in [28] to sequence *g22* (d)-(e); Foreground seeds selected using the audio-visual coherence; Results when the audio-visual term is not used ( $\lambda_C = 0$ ); Our results when both audio-visual and regional terms are considered. In all situations the current speaker is detected.

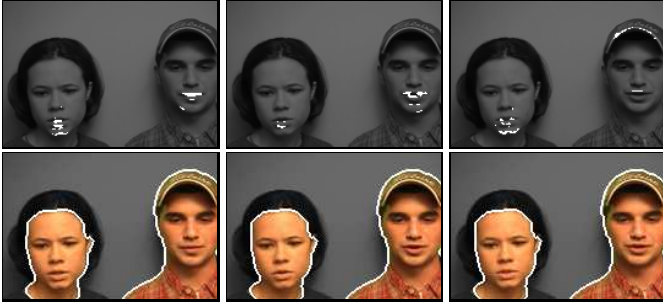


Fig. 8. Results on a fragment of sequence *g21* of CUAVE database in which two persons speak simultaneously. The foreground seeds in these frames are depicted on the top row, while the results are shown in the bottom row.

speaker’s mouth since in this region the audio-visual coherence is high. Therefore, the label of the seeds is efficiently spread and the complete mouth region can be extracted.

A fragment of clip *g21* of CUAVE database [40] in which two persons speak at the same time is used to illustrate our approach’s ability in the extraction of multiple active sources. In some of the frames in Fig. 8 the foreground seeds are mainly situated over the left person, while in other frames most seeds are located over the right person. Since in average the seeds are located over the mouth regions of both speakers most of the time, our approach successfully extracts the faces of the two persons (they are both audio-visual objects).

### B. Results on Entire Sequences

Five sequences containing different types of audio-visual objects, distracting video motion and multiple sources with

different activity patterns compose our dataset (see Table III). Three piano sequences are taken from state-of-the-art audio-visual source localization approaches and depict scenes presenting distracting motion in the camera field of view. *Piano1* [18] features a hand playing a synthesizer (non-stationary sound source) and a wooden rocking horse moving in the background (distracting moving object). The video is sampled at 25 fps with a resolution of  $720 \times 576$  pixels and the audio at 44.1 kHz. For its analysis, the video has been resized to  $240 \times 192$  pixels. In *Piano2* and *Piano3* [41] a hand is playing a piano while a toy car crosses the scene in and a fan is moving in the background, respectively. The original video signals are sampled at 25 fps with a resolution of  $352 \times 288$  pixels and the audio at 48 kHz. The videos have been resized to  $176 \times 144$  pixels for its analysis. *Speakers1* corresponds to a fragment of clip *g14* of CUAVE database [40] in which two persons speak in turns, first the left one and then the right one. The video is sampled at 29.97 fps with a resolution of  $720 \times 480$  pixels and the audio at 44.1 kHz. For its analysis, the video has been resized to  $176 \times 120$  pixels. *Speakers2* is composed by two persons speaking in turns. Unlike in clips from the CUAVE database, the speakers are not situated in front of a green flat background but in a realistic office environment. This movie is recorded with an iSight camera integrated into a MacBook Pro laptop at 25 fps with a resolution of  $640 \times 480$  pixels, and it is resized to  $240 \times 180$  pixels for its analysis. The audio signal is sampled at 44.1 kHz. The propagative procedure in Sec. V is extremely challenged when sources pass from active to inactive or vice versa, since the transfer of the information from one GoF to the next one can be counterproductive.

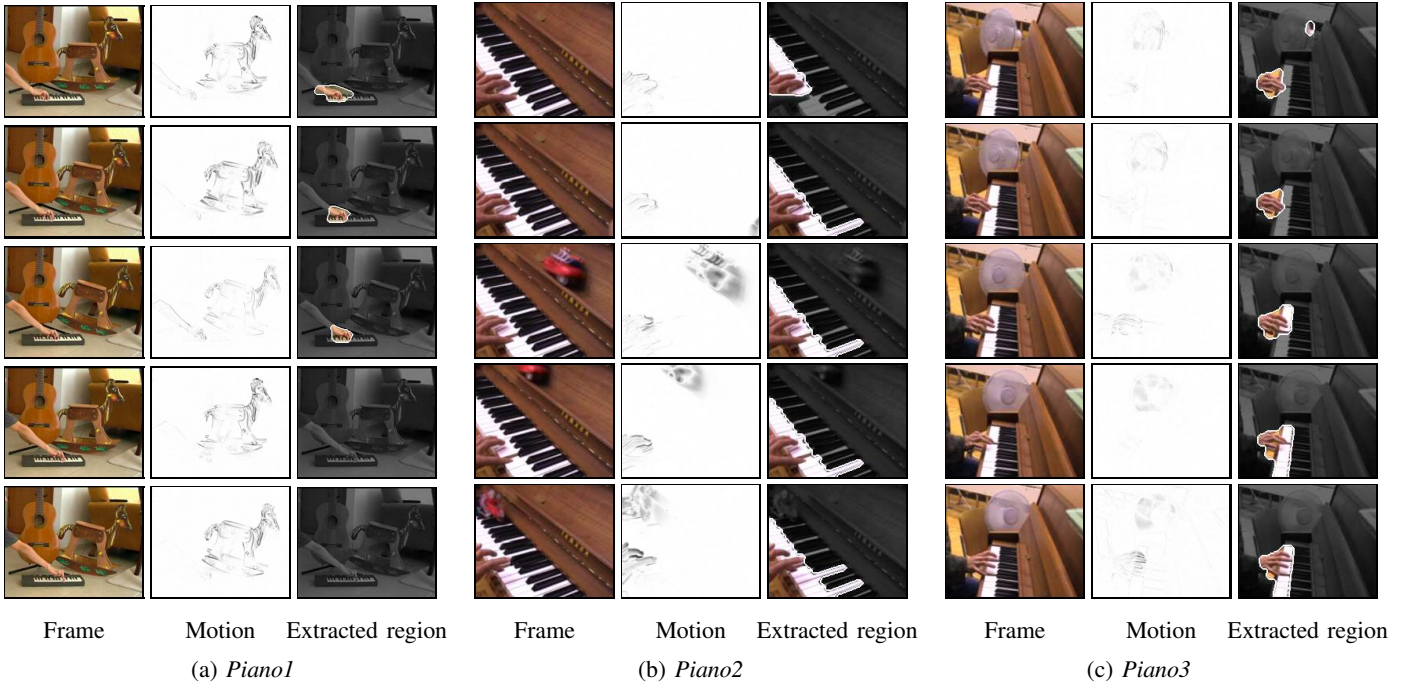


Fig. 9. Audio-visual objects extracted by the proposed method in the presence of distracting motion, which is generated by a rocking horse in *Piano1* (a), a toy car in *Piano2* (b) and a fan in *Piano3* (c).

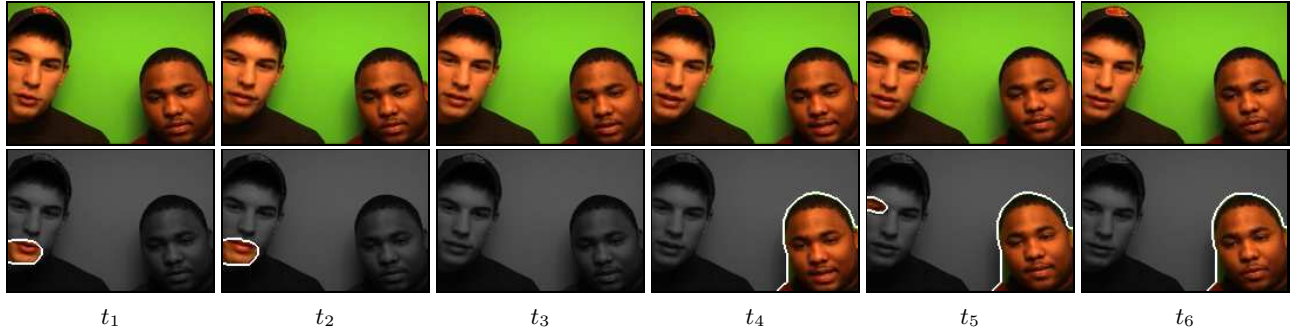


Fig. 10. Results obtained for a fragment of the *Speakers1* sequence where two persons speak in turns. Frame  $t_3$  corresponds to a silence between the periods in which only the left person speaks (frames  $t_1$  to  $t_2$ ) and only the right person speaks (frames  $t_4$  to  $t_6$ ).

Results obtained with our approach when extracting the audio-visual objects from these sequences are shown in Fig. 9, Fig. 10 and Fig. 6(b). The audio-visual object (hand) is extracted for the entire duration of sequences *Piano2* and *Piano3* (Fig. 9(b)-(c)). In contrast, the hand can not be extracted in the last part of clip *Piano1*, since in the final frames (at the bottom in Fig. 9(a)) the motion in the audio-visual object is very small compared to the distracting motion (rocking horse). In this case the foreground audio-visual seeds are divided between both regions and there is not enough concentration of seeds in the hand to allow its extraction. Even if the sequence is so complex, the distracting moving object is not contained at any time in the extracted region. The current speaker is always correctly detected in sequences *Speakers1* and *Speakers2*. However, the entire face region is not extracted for the left speaker in *Speakers1* (frames  $t_1$  and  $t_2$  in Fig. 10). Since the proposed audio-visual segmentation approach is unsupervised we do not have control over the extracted region. Small and sporadic artifacts are extracted in

some cases, as the left person's eye in *Speakers1* (frame  $t_5$  in Fig. 10) or a fragment of the fan in *Piano3* (top frame in Fig. 9(c)). However, small regions extracted during a short period of time can be efficiently eliminated by simply eroding/dilating the segmented region both in space and time.

Videos showing the original sequences and audio-visual objects extracted with the proposed approach are available online at [http://www.eecs.qmul.ac.uk/~llagostera/AVObjectExtraction\\_results.htm](http://www.eecs.qmul.ac.uk/~llagostera/AVObjectExtraction_results.htm).

Table III provides a quantitative analysis of the results in this section. In the computation of *precision* and *recall* measures, a true positive *TP* is defined as a successful extraction of the audio-visual object (i.e. mouth region of the current speaker or hand playing the piano). A false positive *FP* is produced when our method extracts part of the distracting moving object or part of the background. A false negative *FN* occurs when the proposed approach is not able to extract an active audio-visual source. By combining *precision* and *recall* values we can quantify our method's ability in extracting the audio-visual

TABLE III

PRECISION AND RECALL IN EXTRACTING THE AUDIO-VISUAL OBJECT FOR THE ANALYZED SEQUENCES. AVERAGE VALUES ARE COMPUTED BY TAKING INTO ACCOUNT THE NUMBER OF FRAMES OF THE SEQUENCES.

Sequence	Length (s)	Precision (%)	Recall (%)
Piano1	4	100	59
Piano2	3	100	100
Piano3	10	93	100
Speakers1	7	90	87
Speakers2	9	88	96
Average		92	90

object without extracting at the same time distracting moving objects or inactive sources. The proposed method provides a good accuracy in the extraction of the active audio-visual sources in the scene, by leading to high values in *precision* and *recall*. Approximately half of the errors in the audio-visual object extraction (*FP* and *FN*) occur in the transitions between the sources' activity periods (when they pass from inactive to active or vice versa). The other half (53%) of *FN* are the final frames of *Piano1*, in which the hand is not extracted due to the magnitude of the distracting motion. Finally, 54% of *FP* are composed of frames in which a fragment of a video distractor is extracted, i.e. small part of the fan at the beginning of *Piano3* and the inactive speaker's eye in *Speakers1*. Even if these *FP* can be removed with a simple post-processing step penalizing small regions extracted for a short period, the errors in transitions between the sources' activity periods are difficult to eliminate since they result from the division of the signals into GoFs. However, in all experiments the delay between the time in which an audio-visual source becomes active and its extraction is small (less than one second), and the same happens when sources pass from active to inactive. Even if *precision* and *recall* values are already high, we expect our method to improve its performances in both quantities when analyzing sequences representing less challenging situations.

## VII. DISCUSSION

We have presented a novel method that automatically extracts the audio-visual objects in a scene. Video regions presenting high synchrony with the soundtrack are used as the starting point for a graph cut segmentation procedure whose goal is to extract the video modality of the sound source. The knowledge obtained from joint audio-visual processing is used in the selection of the segmentation priors and in the energy function that the graph cuts minimize. A propagative procedure allows the extraction of audio-visual objects in long sequences by ensuring the temporal continuity of the result.

Our approach has been tested in challenging sequences with various types of audio-visual sources, distracting moving objects and multiple sources with different activity patterns. Our definition of the segmentation problem, with both an audio-visual term and a regional term, makes our method more suitable than previous approaches for applications that require

the extraction of complete audio-visual objects. We have demonstrated that our method is able to distinguish between audio-visual objects and distracting moving objects, leading to extracted regions that are stable and evolve according to the changes in the sources activity in a short time delay.

## REFERENCES

- [1] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition," *IEEE Trans. on Multimedia*, vol. 7, no. 3, pp. 495–506, 2005.
- [2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.
- [3] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.
- [4] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (Audiovisual Codebook Dependent Cepstral Normalization)," in *Proc. of Int. Conf. Spoken Language Processing (ICSLP)*, 2002, pp. 1449–1452.
- [5] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2002, pp. 2025–2028.
- [6] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Communication*, vol. 44, no. 1–4, pp. 113–125, 2004.
- [7] R. Dansereau, "Co-channel audiovisual speech separation using spectral matching constraints," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 5, 2004, pp. 645–648.
- [8] S. Rajaram, A. V. Nefian, and T. Huang, "Bayesian separation of audio-visual speech sources," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 5, 2004, pp. 657–660.
- [9] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 96–108, 2007.
- [10] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers, "Video assisted speech source separation," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 5, 2005, pp. 425–428.
- [11] P. Pérez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proc. of the IEEE*, vol. 92, no. 3, pp. 495–513, 2004.
- [12] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 814–820.
- [13] R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio correlation," in *Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME)*, vol. 3, 2000, pp. 1589–1592.
- [14] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proc. of Int. Conf. Image and video retrieval (CIVR)*, vol. 2728, 2003, pp. 488–499.
- [15] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. on Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.
- [16] P. Smaragdis and M. Casey, "Audio/visual independent components," *Proc. of Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, pp. 709–714, 2003.
- [17] G. Monaci, O. Divorra, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Processing*, vol. 86, no. 12, pp. 3534–3548, 2006.
- [18] E. Kidron, Y. Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. on Signal Processing*, vol. 55, no. 4, pp. 1390–1404, 2007.
- [19] Z. Barzelay and Y. Y. Schechner, "Harmony in motion," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [20] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind Audio-Visual Source Separation based on Sparse Redundant Representations," *IEEE Trans. on Multimedia*, vol. 12, no. 5, pp. 358–371, 2010.
- [21] C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann, "Nonnegative cca for audiovisual source separation," in *IEEE Workshop on Machine Learning for Signal Processing*, 2007.
- [22] C. Saraceno and R. Leonardi, "Indexing audiovisual databases through

- joint audio and video processing,” *International Journal of Imaging Systems and Technology*, vol. 9, no. 5, pp. 320–331, 1999.
- [23] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, “Multimedia content processing through cross-modal association,” in *Proc. of ACM Multimedia (MM '03)*, 2003, pp. 604–611.
  - [24] J. Fritsch, M. Kleinhagenbrock, S. Lang, G. A. Fink, and G. Sagerer, “Audiovisual person tracking with a mobile robot,” in *Proc. of Int. Conf. on Intelligent Autonomous Systems*, 2004, pp. 898–906.
  - [25] J. Hershey and J. R. Movellan, “Audio vision: Using audio-visual synchrony to locate sounds,” in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 1999, pp. 813–819.
  - [26] G. Monaci, P. Jost, P. Vanderghenst, B. Mailhe, S. Lesage, and R. Griboval, “Learning Multi-Modal Dictionaries,” *IEEE Trans. on Image Processing*, vol. 16, no. 9, pp. 2272–2283, 2007.
  - [27] Y. Liu and Y. Sato, “Finding Speaker Face Region by Audiovisual Correlation,” in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008*, 2008.
  - [28] Y. Liu and Y. Sato, “Visual localization of non-stationary sound sources,” in *Proc. of ACM Multimedia (MM '09)*, 2009, pp. 513–516.
  - [29] Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images,” in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2001, pp. 105–112.
  - [30] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts,” *Proc. of ACM SIGGRAPH*, vol. 23, no. 3, pp. 309–314, 2004.
  - [31] Y. Li, J. Sun, and H.-Y. Shum, “Video object cut and paste,” *Proc. of ACM SIGGRAPH*, vol. 24, no. 3, pp. 595–600, 2005.
  - [32] X. Bai, J. Wang, D. Simons, and G. Sapiro, “Video snapcut: robust video object cutout using localized classifiers,” in *Proc. of ACM SIGGRAPH*, 2009, pp. 1–11.
  - [33] W. H. Sumby and I. Pollack, “Visual contribution to speech intelligibility in noise,” *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.
  - [34] Q. Summerfield, “Some preliminaries to a comprehensive account of audio-visual speech perception,” in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds. Lawrence Erlbaum Associates, 1987, pp. 3–51.
  - [35] J. Driver, “Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading,” *Nature*, vol. 381, no. 6577, pp. 66–68, May 1996.
  - [36] A. LlagosteraCasanovas and P. Vanderghenst, “Audio-based nonlinear video diffusion,” in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2010, pp. 2486–2489.
  - [37] J. Wang and M. F. Cohen, “Image and video matting: a survey,” *Foundation and Trends in Computer Graphics and Vision*, vol. 3, no. 2, pp. 97–175, 2007.
  - [38] A. LlagosteraCasanovas and P. Vanderghenst, “Unsupervised Extraction of Audio-Visual Objects,” in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2011.
  - [39] BGL, *The boost graph library: user guide and reference manual*. Boston, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.
  - [40] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, “Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus,” *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, p. 1189, Nov. 2002.
  - [41] G. Monaci and P. Vanderghenst, “Audiovisual gestalts,” in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2006.